



Human Population Genetic Variation Data at NCBI

Population-specific allele frequency data to support variant interpretation
<https://www.ncbi.nlm.nih.gov/snp> | <https://api.ncbi.nlm.nih.gov/variation/v0/>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Introduction

NCBI's dbSNP database (www.ncbi.nlm.nih.gov/snp) has been providing researchers with access to information about short human genetic variations since 1999. Consolidation of data submitted by the scientific community has increased the size of this catalog to close to 670 million active human reference SNP records. Understanding the diversity of genetic variations in human populations is of interest in human migration, ancestry and clinical research. Several large research studies have contributed population genetic variation data to dbSNP, including the 1000 Genomes Project, ExAc, GnomAD, TopMed, and others. These datasets are displayed on dbSNP Reference SNP Reports and are available through the E-Utilities and the SPDI API services. Recently NCBI has begun to make use of the wealth of data in the NCBI database of Genotypes and Phenotypes (dbGaP, www.ncbi.nlm.nih.gov/gap), and in collaboration with the NIH Office of the Director, has now made available a subset of aggregated population genetic variation data available in dbSNP. Called the "dbGaP Allele Frequency" project (currently nicknamed ALFA), the benefit of this project is that it adds information from 531.2 million genetic variations based on the self-reported ethnicities of 142,000 subjects. The data have now been added to dbSNP records, which serves as the largest source of population genetic variation data for research purposes.

Viewing Population Data on dbSNP Records

The most convenient way to see population genetic variation data is on individual dbSNP records (right), which list the study-specific allele frequencies at the top (A). Separated by studies, the table in the

"Frequency" tab (B) lists, from left to right, the study, population, sample size, and frequencies for Reference Alleles and Alternate Alleles. Allele frequencies are shown for the studies' global populations (C) and for each of their designated sub-populations (D) if such a breakdown is available. The Download link (E) provides a simple way to save a particular reference SNP's population data in a tab-separated file. The next update will insert a "dbGaP Allele Frequency Project" table (F) at the top of the "Frequency" tab.

API and FTP Access

The E-Utilities' esummary function already provides access to global allele frequencies from included studies. The xtract XML parser from EntrezDirect package makes data extraction straightforward, see details here: <https://www.ncbi.nlm.nih.gov/books/NBK179288/>

NCBI will update the SPDI variation services, the API for SNP, in the near future to support frequency and meta data retrieval. Check this site for update:

<https://api.ncbi.nlm.nih.gov/variation/v0/>

When dbGaP population data for dbSNP is released, dbSNP will make the files available under its FTP directory:

https://ftp.ncbi.nlm.nih.gov/snp//redesign/latest_release/

Reference SNP (rs) Report
Switch to classic site
<https://go.usa.gov/xVMGG>
Current Build 153
Released July 9, 2019

rs328

Organism: *Homo sapiens*
Position: chr8:19962213 (GRCh38.p12)
Alleles: C>G
Variation Type: SNV Single Nucleotide Variation
Frequency: G=0.09216 (23148/251182, GnomAD_exome), G=0.08971 (11265/125568, TOPMED), G=0.09350 (11340/121282, ExAC) (+ 8 more)

Clinical Significance: Reported in ClinVar
Gene: Consequence LPL: Stop Gained
Publications: 117 citations
Genomic View: [See rs on genome](#)

Variant Details
Clinical Significance
Frequency
Aliases
Submissions
History
Publications

Study	Population	Group	Sample Size	Ref Allele	Alt Allele
1000Genomes	Global	Study-wide	5008	C=0.908	G=0.092
1000Genomes	African	Sub	1322	C=0.939	G=0.061
	East Asian	Sub	1008	C=0.878	G=0.122
	Europe	Sub	1006	C=0.870	G=0.130
	South Asian	Sub	978	C=0.91	G=0.09
	American	Sub	694	C=0.94	G=0.06
1000Genomes	Global	Study-wide	614	C=0.87	G=0.13
ExAC	Global	Study-wide	121282	C=0.90650	G=0.09350
ExAC	Europe	Sub	73308	C=0.8998	G=0.1002
ExAC	Asian	Sub	25146	C=0.8971	G=0.1029
ExAC	American	Sub	11532	C=0.9469	G=0.0531
ExAC	African	Sub	10388	C=0.9299	G=0.0701
ExAC	Other	Sub	908	C=0.93	G=0.07

Search:

dbGaP Population Frequency Project Release Version: 20190927154311

Population	Group	Sample Size	Ref Allele	Alt Allele	
gnomAD: Global	Global	11176	C=0.8964	G=0.1036	
gnomAD: Europe	Sub	8134	C=0.896	G=0.104	
gnomAD: All African Ancestry	Sub	676	C=0.91	G=0.09	
gnomAD: 95% Exclusive African Ancestry	Sub	14	C=0.8	G=0.2	
gnomAD: African American	Sub	662	C=0.91	G=0.09	
gnomAD: Asian	Sub	60	C=0.9	G=0.1	
gnomAD: 95% East Asian Ancestry	Sub	28	C=0.9	G=0.1	
gnomAD: South East Asian and Pacific Islanders	Sub	32	C=0.9	G=0.1	
Northern	Sub	0	C=0	G=0	
The Avon Parents	Sub	0	C=0	G=0	
The PAGE	Sub	4	C=0.5	G=0.5	
The PAGE	Sub	2302	C=0.895	G=0.105	
TopMed	Global	Study-wide	125568	C=0.91029	G=0.08971
UK 10K study_Twins	TWIN COHORT	Study-wide	3708	C=0.892	G=0.108

Featured Diverse Population Data Sets

These studies have provided large sets of variations to the dbSNP database for a diverse collection of study subjects.

Project	Description & Links	Subject counts (thousand)	Variants (million)
dbGaP Allele Frequency (ALFA - subject to change)	Culled from well-curated dbGaP populations from Genome Wide Association Studies deposited in dbGaP. https://www.ncbi.nlm.nih.gov/gap/ https://ftp.ncbi.nlm.nih.gov/pub/factsheets/Factsheet_dbGaP.pdf https://ftp.ncbi.nlm.nih.gov/pub/factsheets/FAQ_dbGaP_Data_Request.pdf	142.0	531.2
1000 Genomes	The 1000 Genomes Project aggregated short and structural variation information from the genomes of volunteer subjects with the goal of finding most genetic variants with frequencies of at least 1% in the populations studied. https://www.internationalgenome.org/ https://www.ncbi.nlm.nih.gov/bioproject/PRJEB6930	2.5	84.9
TopMed	The Trans-Omics for Precision Medicine (TOPMed) program consists of data from over 80 different studies with varying designs. https://www.nhlbiwgs.org/ https://www.ncbi.nlm.nih.gov/bioproject/PRJNA400167	6.2	549.4
ExAc	The Exome Aggregation Consortium (ExAC) dataset spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. http://exac.broadinstitute.org/ https://www.ncbi.nlm.nih.gov/bioproject/PRJEB8661	60.7	10.1
gnomAD (Genomes Exomes)	The Genome Aggregation Database (gnomAD) is an extension of ExAc, and aggregated population information from exome and genome sequencing data from a variety of large-scale sequencing projects. https://gnomad.broadinstitute.org/ https://www.ncbi.nlm.nih.gov/bioproject/PRJNA398795 (genomes) https://www.ncbi.nlm.nih.gov/bioproject/PRJNA398795 (exomes)	142.0	228.7
GO Exome Sequencing Project	The NHLBI GO Exome Sequencing Project (ESP) studied factors contributing to heart, lung and blood disorders with next-generation sequencing of protein-coding regions for diverse, richly-phenotyped populations. https://evs.gs.washington.edu/EVS/ https://www.ncbi.nlm.nih.gov/bioproject/PRJNA192955	6.5	1.4
The PAGE Study	The PAGE study imputed genotyped over 50,000 samples to the 1000 Genomes panel, and included sequencing of 1,000 additional samples representative of 21 populations in the Americas. http://www.pagestudy.org/ https://www.ncbi.nlm.nih.gov/bioproject/PRJNA168052 https://go.usa.gov/xVMSt (dbGaP Study)	39.4	1.3

Featured Regional Population and Cohort Data Sets

The table below summarizes projects with limited scope that focus on regional populations of interest.

Project	Description & Links	Subject counts (thousand)	Variants (million)
The Avon Longitudinal Study of Parents and Children	Genotypes of 14,500 families in the Bristol area. http://www.bristol.ac.uk/alspac/ https://www.ncbi.nlm.nih.gov/bioproject/PRJEB7217	3.9	46.6
UK 10K Study – Twins	14,274 Twins from 76 Studies https://twinsuk.ac.uk/ https://www.ncbi.nlm.nih.gov/bioproject/PRJEB7218	3.7	46.6
Vietnamese Genetic Variation Database	305 genomes and exomes of healthy Kinh Vietnamese. Found more than 99.3% of common variants with a frequency of above 1%. https://genomes.vn/ https://www.ncbi.nlm.nih.gov/bioproject/PRJNA515199	3.5	24.8
Genetic variation in the Estonian population	Genome- and pharmacome-wide associations for 2240 Estonian Biobank participants. https://www.nature.com/articles/s41431-018-0300-6 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA489787	2.2	31.7

General Help, Comments, and Feedback

Please refer to NCBI Factsheets collection for quick start on NCBI resources: http://bit.ly/ncbi_factsheets

For questions and comments, please write to info@ncbi.nlm.nih.gov